WEB SEARCH ENGINE FOR PAGE RANKING ALGORIHTM

AP18110010467, AP18110010468, AP18110010491, AP18110010518

Abstract

A Web search engine is a specialized computer server that searches for Web information. User query search results are often returned as a list of URLs. Some search engines search and retrieve data from social media or open indexes. Web Search Engine is a web crawling software that identifies you and provides information based on user search query. Some search engines go through it and extract information from various open-source databases. Often, search engines offer real-time results based on the algorithm they use to retrieve and analyse data. Search engines play an important role in Web success, search engines help any Internet user quickly find relevant information.

They are essentially very large data mining applications. Various data mining techniques are used in all aspects of search engines, ranging from *crawling*, indexing (e.g., you select the pages that will be is displayed and decides how much the index should be constructed), and searching (e.g., decide the process for ranking the pages, adding required advertisements and how the search results can be personalized).

The algorithms discussed in this report are the page rank algorithm and the HITS algorithm.

Introduction

Users use search engines for most of their queries but only like the results found on the first page and then 2-3% of users go to other pages (excluding search engines), Now imagine if the organization page is 2-3 or 4 page then the business that can be generated from that page has very little change to retrieve and the user will select the page that comes on the 1st page.

Billions of web pages are indexed daily by a search engine. Millions of searches are happening per day. Most visitors visit the website with

hitting the links found in search engines and believing that companies found in high results are the best product in the product

service and its category. These indicators make it very clear that if an organization wants to move forward in its sales then they should focus on getting their page widely available on search engines.

There are various page ranking algorithms proposed till date to improvising the search engine results. Firstly Page ranking algorithm proposed by Sergey Brin and Larry Page was Page Rank algorithm. It ranks the page by using the number of links attached to it. But Page rank has the limitation that because of dividing the rank equally among the pages results in low rank values which leads to rank pages with less relevant content to be listed on top of the page urls. Second is HITS Hypertext Induced Topic Search (HITS)or link analysis algorithm. The basic version of HITS algorithm calculates page rank based on Hubs and authorities.

Advantages:
It gives the best search result to the user by using the latest tools, algorithms and technologies. The features used by it are search engine optimization, web crawling, indexing, page ranking etc. to extract the exact and fast result.

LITERARTURE SURVEY:

Olston and Panday [1] crawled 10,000 random samples of URLs and 10,000 pages sampled from the Open Directory every second days for several months. Their analysis measured both change frequency and information long evity are the average lifetime of a shingle, and found only a moderate correlation between the two. They introduce new crawl policies that are aware to information longevity

P. Ravi Kumar [2] Put a light on we bsite optimization, its challenges as well as website optimization classes. The author provides a detail report on WSO “Onsite” and “Offsite” ranking factors, efficient methodologies and techniques for onsite and offsite ranking factors. There is also a list of black hat techniques that can potentially reduce the ranking of a website and even can remove the entry from the index of a search engine.

Bahador Saket and Farnaz Behrang [3] presented a technique to determine correctly the quality of links that have not been retrieved so far but a link is accessible to them. For this reason, author apply an algorithm like an AntNet routing algorithm.

Eytan Adar et.al [4] described algorithms, analyze, and models for characterizing the evolution of Web content. Proposed analysis gives insight into how Web content changes on a finer grain than previous study, both in terms of the time intervals studied and the detail of change analyzed.

Meenakshi Bansal [5] Emphasizes that website optimization is about making important modifications to almost concerned sections of the website.Though it is viewed individually , some of the change might seem like gradual incremental improvements. SEO is a process which requires considerable time. This work on on-page optimization includes actual code merged with various languages, keyword placement and keyword density.The search engine employs combination of automated algorithms, manually edited directories and advertisements to generate results for users' queries.

PROPOSED SYSTEMS:

The proposed system uses the latest algorithms such as search engine optimization technique, page ranking indexing and web crawling. This will provide the best search results to the users. The unique and distinct search result is displayed by the proposed search engine to the user's query. A number of web links are indexed whenever a keyword is searched by the user using this Search Engine application. The analysis is perform for the data in the indexed links. This task is achieved by page rank algorithm. Finally, the web page containing the top match to the key word is showed to the user in the first link of the result page.

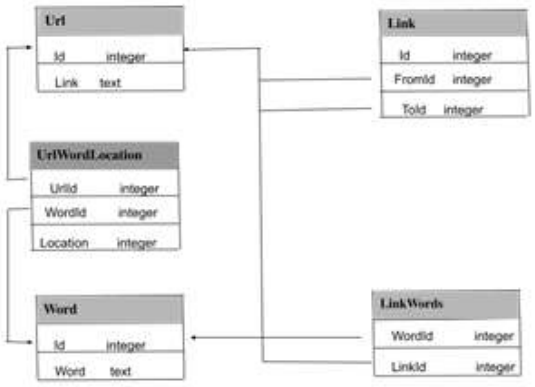
Modules

Search: User search for the particular query in the search engine This user query is saved for further processing.

Indexing: After a user searches for the particular keyword, database matching those keywords is identified. Then number of web links relevant to this query are indexed.

Data Base tables :

Figure



we created 5 database tables.

1. Url - Stores each URL and assigns an id
2. Word - Stores each word from a url's HTML page

3. UrlWordLocation - Location or index of each word inside a url's

HTML page thats why foreign keys are url id and word id
4. Link - Stores from which page or page url it came from and which page url it is pointing to.
5. LinkWords - Additional table might not be required but stores words in the url's text.

Analysing:
Then to retrieve the best possible match, this search engine application uses page ranking, web crawling and search engine optimization.

Result Retrieval:
After analysing, the best matching result is retrieved and it is displayed in the user interface window to the user.

Software Requirements
Windows OS
Python IDE – Jupyter Notebook
SQLite
Hardware Requirements
Hard Disk – 500 GB or Above
RAM required – 4 GB or Above
Processor – Core i3 or Above

ALGORITHMS:

1. Crawler:

- Google founders Sergey Brin and Lawrence Page, in their seminal paper, identified the Web crawler as the most sophisticated yet fragile component of a search engine [9]

- Web crawlers, also called as spiders/robots. Since information on the Web is scattered among billions of pages served by millions of servers around the globe,users who browse the Web can follow hyperlinks to access information,virtually moving from one page to the next page.

- crawler can crawl all web pages of a website and index them. Crawler takes URL as input,index it and fetches all links on that page, crawler then moves on to previously fetched links, index them and find more links on each page and this process goes on until all URLs are indexed.
- Crawler is necessary for search engines to index web. In this project, we will learn how we can create a crawler and store page content into DB, and how to make a simple searchengine.We will use some packages like url lib, Beautiful Soup and sqlite3.

- The most extensive use of crawlers is, however, in support of search engines. In fact, crawlers are the main consumers of Internet bandwidth.Well liked search engines such as Google, Yahoo! and MSN run very efficient universal crawlers designed to gather all pages irrespective of their content.

Fig 1.1 shows the flow of a basic sequential crawler. Such a crawler
fetches one page at a time, making inefficient use of its resources.

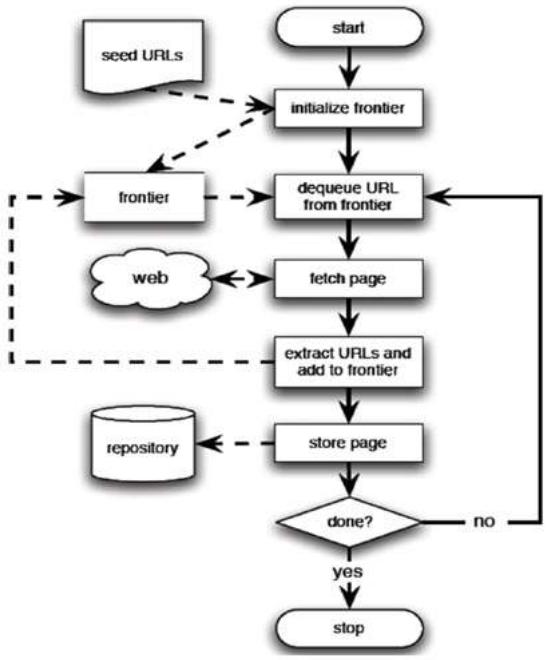


Fig 1.1 figure

- A crawler is, in essence, a graph search algorithm. The Web can be shown as a big graph with pages as its nodes and hyperlinks as its edges.
- A crawler begins from a few of the nodes (seeds) and then follows the edges to reach other nodes. The process of fetching a page and extracting the links within it is compitable to expanding a node in graph search.

Basic crawler algorithm:

In each iteration of its main loop, the crawler picks the next URL from the frontier, fetches the page corresponding to the URL through HTTP, parses the retrieved
page to extract its URLs, adds newly discovered URLs to the frontier, and stores the page (or other extracted information, possibly index terms) in a local disk repository.

The crawling process may be terminated when a total number of pages have been crawled.

Inverted index:

An inverted index is an index data structure storing a mapping from content, such as words or numbers, to its locations in a document or a set of documents. In simple words, it is a hash map like data structure that directs you from a word to a document or a web page.

Similar to an index of a book, a search engine also extracts and builds a

cata-log of all the words that appear on each web page and the number of times it appears on that page [6]. The parser can extract the relevant information from a web page by excluding certain common words (such as a, an, the - also known as stop words). Indexes are updated periodically as new content is crawled. Some indexes help create a dictionary (lexicon) of all words that are available for searching.

Methods of Indexing:

1) Full-Text Indexing As its name suggets, full-text indexing is where every word on the page is put into a database for searching. Full text indexing will help to find every example of a reference to a specific name or terminology.

2) Human Indexing Yahoo and some of Magellan are some of the examples of human indexing. In the Keyword indexing, all of the work was done by a computer program called a "spider" or a "robot"[7].

Steps to build Inverted index are:

a. Fetch the document and gather all the words.

b. Reading the document and extracting and tokenizing all of the text create new entry for each token then add a index to the database table.

c. Repeat above steps for all documents

d. Indexing is slow as it first checks that word is present or not.

e. Searching is a quick process.

Real life example: Index at the back of the book

Query indexing:

Searching for relevant documents in the inverted index consist of

- Vocabulary/keyword search: This step finds each query term in the inverted index table, which gives the inverted list of each term.

Keyword:

A keyword is a word or phrase an Internet user will enter into a search engine when trying to locate something, i.e., a product or information. For example, a website selling herbal tea will list keywords such as, “herbal,” “tea,” and “tea bags,” etc.

PAGE RANKING

In the early 90s, the first search engines used text based ranking systems to decide which pages are most relevant to a given query.

Main idea: We say a page is important when it is pointed by other main pages.

- One of the most known and influential algorithms for

computing the relevance of web pages is the Page Rank algorithm used by the Google search engine.

- PageRank was presented by Sergey Brin and Larry Page at the Seventh International World Wide Web Conference (WWW7) in April, 1998.
- Based on the algorithm, they built the search engine Google.
- It is a query-independent evaluation of Web pages.
- Page Rank interprets a hyperlink from page x to page y as a vote. (by page x, for page y)
- It analyses the page that casts the vote.
- Votes casted by pages that are themselves "important" weigh more heavily and help to make other pages more "important."
- PageRank is a limited version of Web pages in the sense that PageRank value can be removed when the off-line is cut and can be removed for private inquiries.
- In-links of page i: This is because linked links are able to touch anything. Usully, hyperlinks added to the site and were not found.
- Out page i links: These are the links that affect other things than i. Usually, links affect site location and not found.
- PageRank used for special networks availability. value of each page can will be reset (rank).

From the perspective of prestige, the following are used to derive the PageRank algorithm.

1. A hyperlink from a page pointing to another page is an implicit conveyance of authority to the target page. Therefore the more in-links that a page i receives, the more prestige the page i have.

2. Pages that point to page i also have their own prestige scores. A page with a higher prestige score pointing to i is more important than a page with a lower prestige score pointing to i. In other words, a page is important if it is pointed to by other important pages.

- According to prestige-rank in social networks, the importance of page i (i's PageRank score) is determined by summing up Plagiarizedthe PageRank scores of all pages that point to i. Since a page may point to many other pages, its prestige score should be shared among all the pages that it points to.
- To formulate the above ideas, we treat the Web as a bi-directed graph G = (V, E),where V is the set of vertices or nodes, i.e., the set of all pages, and E is the set of directed edges in the graph, i.e., hyperlinks.

Let the total number of pages on the Web be n (i.e., n = |V|).

The PageRank score of the page i (denoted by P(i)) is defined by:

$$P(i) = \sum_{(j,i)\in E} \frac{P(j)}{O_j},$$

where Oj is the number of out-links of page j.

Hits Algorithm:

HITS stands for Hyper Text Induced Topic title search, static status algorithm, HITS Search query depends. when a user releases a search query, the first HITS expands the list of related pages are retrieved by the search engine and extract two levels of expansion page set, position of authority and hub level.

An authority is a page with many links. a great idea that a page can have authoritative content on another topic so most people trust and interact with it.

The hub is a page with many external links. This page acts as an information editor at a specific topic and point to many pages of good authority in the article. When the user comes on this hub page / we will find many helpful links that lead us to great content pages at topic.

The core concept of HITS is a good hub for many good and good executives authority is shown by many beautiful places. Authorities and hubs work together strengthening relationships.

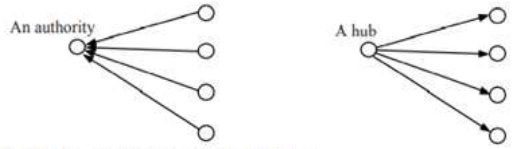


Fig. 7.8. An authority page and a hub page

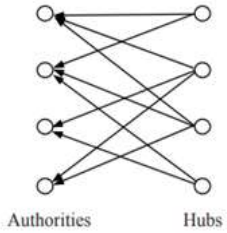


Fig. 7.9. A densely linked set of authorities and hubs

This Algorithm stands for Hyperlink-Induced Topic Search (HITS) Algorithm. This algorithm is given by Jon Kleinberg. This algorithm is also as called Link Analysis algorithm. This is used for ranking the web page focusing on Hubs and Authority. When a user issue some search query HITS algorithm expands list of web pages returned by search engines in response to a user search query. Hits algorithm is query dependent. The famous Twitter web site uses the HITS style algorithm.

HUB: Hub represents a page that points to authorities.
Authority: Authority represents as a source of important information.

Steps of HITS Algorithm [8]:

Step 1: Enter web adjacency matrix pages.

Step 2: Enter the frequency of the various parameters (No for different clicks, and keyword).

Step 3: Rate each web page Hubs and Authorities.

Step 4: Arrange all the attributes to be rated on each web page and then get a per page per page.

Step 5: Then add the parameter values to the fixed volume.

Step 6: Sort the page rank according to the integrated rankings associated with both Hub Values and the Value Authority page of the website.

Step 7: Exit.

HITS-Iterate(G)
$a_0 \leftarrow h_0 \leftarrow (1, 1, \ldots, 1)$;
$k \leftarrow 1$
Repeat
$a_k \leftarrow L^T L a_{k-1}$;
$h_k \leftarrow L L^T h_{k-1}$;
$a_k \leftarrow a_k / ||a_k||_1$; // normalization
$h_k \leftarrow h_k / ||h_k||_1$; // normalization
$k \leftarrow k + 1$;
until $||a_k - a_{k-1}||_1 < \varepsilon_a$ and $||h_k - h_{k-1}||_1 < \varepsilon_h$;
return a_k and h_k

The iteration ends after the 1 norm of the residual of vectors are less than some thresholds a and h. The pages with large authority and hub scores are better authorities and hubs respectively.

HITS will select a few top ranked pages as authorities and hubs and then return to the user .

CONCLUSION:

Search Engine is designed for getting relevant results. The primary goal is to provide high quality search results over a rapidly growing World Wide Web.

Search Engine Project in Python is an excellent web searching platform. It gives fast, accurate and efficient result with the implementation of modern searching algorithms. The fast growing use of internet confirms the good future and scope of the proposed project.
We have seen how to create a simple Crawler which crawls a website and index all its pages as well a simple search engine is implemented which can search keywords.

Future Work:

The precision of web search engine implemented is pretty decent and acceptable, given that the fact that it was developed from the scratch.

But performance of the search engine can be improved to next level with different ranking techniques for a web page.
--- Introduction of new ranking techniques from the current ranking algorithms.

- Implement leadership page ranking
- A GUI application can be made as front end for the search engine.

References

[1] Olston, C. and Pandey, S. Recrawl scheduling based on information longevity. WWW '08, 437-446, 2008.

[2]. P. Ravi Kumar, Ashutosh K. Singh, "Efficient Mehodologies to optimize website for Link Structure based Search Engines", 978-1-4673-6126-2/2013, IEEE

[3] Bahador Saket and Farnaz Behrang "A New Crawling Method Based on AntNet Genetic and Routing Algorithms", International Symposium on Computing, Communication, and Control, pp. 350-355, IACSIT Press, Singapore, 2011

[4] Eytan Adar, Jaime Teevan , Susan T. Dumais and Jonathan L. Elsas "The Web Changes Everything: Understanding the Dynamics of Web Content", ACM 2009

[5] Meenakshi Bansal, Deepak Sharma, ,"Improving webpage visibility in Search Engines by enhancing keyword Density using improved On-Page Optimization technique", IJCSIT, 5347-5352, ISSN0975-9646, 2015

[6] C. W. Cleverdon. The Cranfield tests on index language devices. In Aslib Proceedings, volume 19, pages 173-192, 1967. (Reprinted in Readings in Information Retrieval, K. Spärck- Jones and P. Willett, editors, Morgan Kaufmann, 1997)

[7] Mike Barus. “Link Exchange and One Way Links Using Web Directories,” February 2009.

[8] P.S. Nisha, “A Review Paper on SEO based Ranking of Web Documents”, IJARCSSE, vol. 4, Issue 7, (2014).

[9].Brin, S. and P. Lawrence. The anatomy of a large-scale hypertextual web search engine. Computer Networks, 1998, 30(1-7): p. 107-117.

Group members:

CSE-H

AP18110010467-V. Sai Nayani
AP1811010468- B. Rajya Lakshmi
AP18110010491- T. Vani
AP18110010518- V. Mounika